Validating new discoveries in Sports Medicine – we need FAIR play beyond p-values

British Journal of
Sports Medicine

Validating new discoveries in Sports Medicine – we need FAIR play beyond p-values

Journal:	*British Journal of Sports Medicine*
Manuscript ID	bjsports-2019-101797.R2
Article Type:	Editorial
Date Submitted by the Author:	n/a
Complete List of Authors:	Bleakley, Chris; University of Ulster at Jordanstown; High Point University Smoliga, James; High Point University
Keywords:	Validation, Statistical review, Research, Methodology

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Reviewer: 1

Comments to the Author
I think the authors for their consideration of the previous comments and how they've addressed the issues, and am happy to see this go through with one thing to be resolved. Sorry to be that guy here, but I think there's a misunderstanding on MDC and MCID.
I think you have to consider these as additive, not independent as you've inferred from your figure. In essence, if the patient thinks 2cm matters, and there's a 1cm measurement error, then you have to see 3cm before you can be confident that your observed effect is meaningful to the patient (their 2cm, plus the possibility of a 1cm error). Of course no-one ever does this, but this is important for the reasons you say. As you're aware, it's pretty rare for patients to ever be consulted in this stuff, so the idea of framing results in a context that's important to patients is the key point here, not the technicalities of how to do it.
Rod Whiteley

Many thanks for your input again on this. We have corrected the graph to better highlight that MDC and MCID are additive (rather than independent). We have also included the following –

Although study B[7] reports a larger average effect, most of the observed changes do not reach the threshold for clinical importance (MDC+MCID) and are unlikely to be meaningful to patients.

Reviewer: 2

Comments to the Author
I commend the authors for addressing all of my concerns skillfully and satisfactorily. This editorial should aid clinicians and academics alike. This work provides a useful systematic framework to aid in interpreting new sports medicine findings, in a conscious and vigilant fashion.

Many thanks

Reviewer: 3

Comments to the Author
Please see attached file

While some of my concerns from the first round of reviews have been addressed (mandatory standard, wording on a priori registration, harmonization vs replication), many more have not been sufficiently addressed. My main concern is that this editorial does not bring much new to the table, except for the collection of four individually important (and quite well known) topics under an acronym. By bringing these topics together, we should aim for something more than to tell readers that these topics should be considered. I lack a sense of how to operationalize FAIR, to make it useful. While the evidence is clear for the individual topics that make up FAIR, there is no evidence- base for the usefulness of FAIR itself.

Many thanks for reviewing this. We agree that many components of FAIR are well known, but we have provided evidence within the editorial that these components continue to be overlooked by both clinicians and researchers. We have amended the last paragraph to acknowledge FAIR as a preliminary concept.

FAIR is presented as a preliminary concept to help clinicians disentangle true positive from potentially false positive claims within sports medicine.

I add here just a few specific comments:

1. In the section on False Positive Risk, I find the example to be confusing for the topic of this section. I'm not sure how this can be useful for readers to assess the elevated risk of false discovery. More information is needed to show readers how to use this.

We have amended the second paragraph of this section as below:

FPR calculation is underpinned by Bayes' Theorem, whereby information from two sources (the prior probability of treatment success AND the data from the experiment), are combined to provide a "posterior probability" of treatment success. When appraising experimental research, we can reverse this logic using the data to estimate the prior probability of treatment success; whilst being cognizant that a neutral prior (a 50:50 chance of treatment success), is perhaps the largest that can be legitimately assumed.[2] For example, an experimental study (n=20 per group) reporting a large effect size (1.1) and a p-value of 0.049, corresponds to a prior probability of 97% - if we assume a FPR of 5%. Such an inflated prior suggests the experiment was potentially unnecessary (as the researcher was almost certain of treatment success at the study's inception), OR that the FPR exceeds the set threshold (eg. 5%) and there is elevated risk of false discovery.

2. In my original review, I asked why the authors do not mention confidence intervals. The section on Clinical Importance have been updated; however, the focus is on effect sizes. First, why effect sizes and not **effect measures**? You even criticize this yourselves ("as they are standard scores [...] their clinical context is limited"). Effect estimates of effect measures, with confidence intervals, convey clinical context, and clinically relevant interpretations. Second, the description of confidence intervals as providing "potential range of an effect" is not entirely fair, as confidence intervals reflect the precision in the estimates, which is an important piece of information not reflected in P-values.

Many thanks – we now amended this section:

P-values do little to indicate the clinical importance of observed treatment effects. Effect measures are more intuitive, but standard scores (eg. standardized mean difference) don't provide immediate clinical context. Therefore, legitimate clinical importance can only be determined by framing the difference in means (+ confidence intervals) with relevant Minimal Detectable Change (MDC) and Minimal Clinical Important Difference (MCID) thresholds. MDC represents 'the amount of change (in the outcome) that must be observed before it is considered above the bounds of measurement error'; and MCID represents 'the smallest change (in the outcome) that would be important to patients'. These thresholds are commonly overlooked, and a 2018 audit found that just 7% of orthopaedic researchers referred to MCID when determining treatment effects.5

3. I do feel I have to return to the example in Figure 1. There is something funny going on in this example. Yes, there is not always agreement between a 95% CI and a 5% significance test; however, in this example, we have the mean of a continuous variable, for which you would need to be very creative in your choice of CI and test to achieve a 95% CI as that for A in Figure 1 and a $P < 0.05$ for the corresponding test. The lower limit of the CI is approx. -1 cm, which in this case is quite far from the null effect of 0. Maybe this is a small point, but I'm left with the impression that there is something incorrect here.

Many thanks – we have double checked the data from these studies and the error bars remain unchanged. However, we agree that such a large overlap may cause some confusion to readers, or may even be the result of a reporting error on the original publications. Therefore, we selected different data and amended Figure 1.

Validating new discoveries in Sports Medicine – we need FAIR play beyond p-values

Bleakley C [1,2] (0000-0001-9032-9691), Smoliga JM[2] (0000-0002-1895-5687)

1School of Health Science, Ulster University, Shore Road, Newtownabbey, Northern Ireland
2Department of Physical Therapy, Congdon School of Health Science, High Point University, 1 University Parkway, High Point, NC 27260

Chris Bleakley associate professor James Smoliga professor

Corresponding author
Chris Bleakley
Ulster University
Newtownabbey
Northern Ireland
c.bleakley@ulster.ac.uk

Word count: 792

There is concern that a large proportion of scientific research is based on false positive, non-replicable conclusions.[1] As most experimental research in Sports Medicine is based on frequentist reasoning, p-values have been at the center of knowledge claims and new discoveries within this field. But many researchers and clinicians are unable to define or accurately interpret p-values. Common misconceptions are that p-values represent 'the probability that the null hypothesis is true' or 'the probability that the hypothesis being tested is true.'[2] In effect, p-values only quantify the chances of getting the observed data (on the assumption that the null hypothesis is true), and therefore cannot exclusively inform clinical decision making. This editorial presents FAIR: a 4-item approach to help validate new discovery in sports medicine.

1. **F**alse Positive Risk (FPR)

FPR is "the probability of observing a statistically significant p-value and declaring that an effect is real, when it is not."[2] Crucially, a study's FPR can be high, even when the corresponding p-values are low. In a systematic audit of high quality randomized controlled trials in sports physiotherapy, 18% of 'statistically significant' findings had a 50% chance of false discovery (claiming a treatment effect is real when it isn't).[3]

FPR calculation is underpinned by Bayes' Theorem, whereby information from two sources (the prior probability of treatment success AND the data from the experiment), are combined to provide a "posterior probability" of treatment success. When appraising experimental research, we can reverse this logic using the data to estimate the prior probability of treatment success; whilst being cognizant that a neutral prior (a 50:50 chance of treatment success), is perhaps the largest that can be legitimately assumed.[2] For example, an experimental study (n=20 per group) reporting a large effect size (1.1) and a p-value of 0.049, corresponds to a prior probability of 97% - if we assume a FPR of 5%. Such an inflated prior suggests the experiment was potentially unnecessary (as the researcher was almost certain of treatment success at the study's inception), OR that the FPR exceeds the set threshold (eg. 5%) and there is elevated risk of false discovery.

2. **A** priori registration

Currently only 1 in 3 RCTs in sports physiotherapy are prospectively registered.[3] *A priori* registration of clinical trials ensures that key study details, including primary outcomes, are made public prior to analysis. Unregistered trials carry a higher risk of false discovery, due to unplanned multiple testing, selected reporting and confirmation bias. Registration can help to control the 'degrees of freedom' a researcher has when making small but important decisions regarding data analysis and reporting.[4] The corollary is that positive conclusions from prospectively registered RCTs should hold most weight; with positive findings from unregistered studies considered as exploratory or even hypothesis generating.

3. Clinical **I**mportance

P-values do little to indicate the clinical importance of observed treatment effects. Effect measures are more intuitive, but standard scores (eg. standardized mean difference) don't provide immediate clinical context. Therefore, legitimate clinical importance can only be determined by framing the difference in means (± confidence intervals) with relevant Minimal Detectable Change (MDC) and Minimal Clinical Important Difference (MCID) thresholds. MDC represents 'the amount of change (in the outcome) that must be observed before it is considered above the bounds of measurement error'; and MCID represents 'the smallest change (in the outcome) that would be important to patients'. These thresholds are commonly overlooked, and a 2018 audit found that just 7% of orthopaedic researchers referred to MCID when determining treatment effects.[5]

Figure 1 shows data from two experimental studies,[6 7] each reporting statistically significant changes in ankle dorsiflexion post intervention ($p<0.05$). Despite this, the treatment effects observed in study A[6] cannot be differentiated from measurement error. Although study B[7] reports a larger average effect, most of the observed changes do not reach the threshold for clinical importance (MDC+MCID) and are unlikely to be meaningful to patients.

Insert Figure 1

Figure 1 footnote:
Dots (and whiskers) represent mean change scores (95% CIs)
Ankle dorsiflexion measured through a weight bearing lunge test (cm).; MDC = 1.9cm; MCID = 2cm

4. Replication

The replication crisis is a ubiquitous and complex problem across all of science. Sports medicine has been slower to react compared to other fields of medicine; currently, the volume of research in this field which has been successfully corroborated through replication is unclear. FAIR reminds clinicians and researchers that independent replication underpins scientific discovery; and that it is presumptuous to conclude treatment effectiveness based on a single significant result.

Summary

Time restraints and lack of training are cited as common barriers preventing clinicians from fully engaging in the evidence base. P-value thresholds (is $p<0.05$?) offer a fast but ultimately limited method for determining clinical effectiveness. Although there are many other aspects of trial design and reporting that can increase the risk of false discovery; FAIR is presented as a preliminary concept to help clinicians disentangle true positive from potentially false positive claims within sports medicine.

Competing interests: no competing interests for any author
Contributorship: Both authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. CB and JS were involved in the concept, design and writing. All authors were involved in final submission and revision of the manuscript.
Acknowledgements: none
Funding info: none
Ethical approval information: not applicable
Data sharing statement: not applicable

References

1. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2(8):e124. doi: 10.1371/journal.pmed.0020124
2. Colquhoun D. The False Positive Risk: A Proposal Concerning What to Do About *p*-Values. *The American Statistician* 2019;73:192-201.
3. Bleakley CM, Reijgers J, Smoliga JM. Many high quality RCTs in sports physical therapy are making false positive claims of treatment effect: a systematic survey. *J Orthop Sport Physio, in press*
4. Forstmeier W, Wagenmakers EJ, Parker TH. Detecting and avoiding likely false-positive findings - a practical guide. *Biol Rev Camb Philos Soc* 2017;92(4):1941-68. doi: 10.1111/brv.12315
5. Copay AG, Eyberg B, Chung AS, et al. Minimum Clinically Important Difference: Current Trends in the Orthopaedic Literature, Part II: Lower Extremity: A Systematic Review. *JBJS Rev* 2018;6(9):e2. doi: 10.2106/JBJS.RVW.17.00160
6. Driller M, Mackay K, Mills B, et al. Tissue flossing on ankle range of motion, jump and sprint performance: A follow-up study. *Phys Ther Sport* 2017;28:29-33. doi:10.1016/j.ptsp.2017.08.081
7. Marrón-Gómez D, Rodríguez-Fernández ÁL, Martín-Urrialde JA. The effect of two mobilization techniques on dorsiflexion in people with chronic ankle instability. Phys Ther Sport. 2015 Feb;16(1):10-5. doi: 10.1016/j.ptsp.2014.02.001

Validating new discoveries in Sports Medicine – we need FAIR play beyond p-values

Bleakley C [1,2] (0000-0001-9032-9691), Smoliga JM[2] (0000-0002-1895-5687)

1School of Health Science, Ulster University, Shore Road, Newtownabbey, Northern Ireland
2Department of Physical Therapy, Congdon School of Health Science, High Point University, 1 University Parkway, High Point, NC 27260

Chris Bleakley associate professor James Smoliga professor

Corresponding author
Chris Bleakley
Ulster University
Newtownabbey
Northern Ireland
c.bleakley@ulster.ac.uk

Word count: 785

There is concern that a large proportion of scientific research is based on false positive, non-replicable conclusions.[1] As most experimental research in Sports Medicine is based on frequentist reasoning, p-values have been at the center of knowledge claims and new discoveries within this field. But many researchers and clinicians are unable to define or accurately interpret p-values. Common misconceptions are that p-values represent 'the probability that the null hypothesis is true' or 'the probability that the hypothesis being tested is true.'[2] In effect, p-values only quantify the chances of getting the observed data (on the assumption that the null hypothesis is true), and therefore cannot exclusively inform clinical decision making. This editorial presents FAIR: a 4-item approach to help validate new discovery in sports medicine.

1. **F**alse Positive Risk (FPR)

FPR is "the probability of observing a statistically significant p-value and declaring that an effect is real, when it is not."[2] Crucially, a study's FPR can be high, even when the corresponding p-values are low. In a systematic audit of high quality randomized controlled trials in sports physiotherapy, 18% of 'statistically significant' findings had a 50% chance of false discovery (claiming a treatment effect is real when it isn't).[3]

FPR calculation is underpinned by Bayes' Theorem, whereby information from two sources (the prior probability of treatment success AND the data from the experiment), are combined to provide a "posterior probability" of treatment success. When appraising experimental research, we can reverse this logic using the data to estimate the prior probability of treatment success; whilst being cognizant that a neutral prior (a 50:50 chance of treatment success), is perhaps the largest that can be legitimately assumed.[2] For example, an experimental study (n=20 per group) reporting a large effect size (1.1) and a p-value of 0.049, corresponds to a prior probability of 97% - if we assume a FPR of 5%. Such an inflated prior suggests the experiment was potentially unnecessary (as the researcher was almost certain of treatment success at the study's inception), OR that the FPR exceeds the set threshold (eg. 5%) and there is elevated risk of false discovery.

2. **A** priori registration

Currently only 1 in 3 RCTs in sports physiotherapy are prospectively registered.[3] *A priori* registration of clinical trials ensures that key study details, including primary outcomes, are made public prior to analysis. Unregistered trials carry a higher risk of false discovery, due to unplanned multiple testing, selected reporting and confirmation bias. Registration can help to control the 'degrees of freedom' a researcher has when making small but important decisions regarding data analysis and reporting.[4] The corollary is that positive conclusions from prospectively registered RCTs should hold most weight; with positive findings from unregistered studies considered as exploratory or even hypothesis generating.

3. Clinical **I**mportance

P-values do little to indicate the clinical importance of observed treatment effects. Effect measures are more intuitive, but standard scores (eg. standardized mean difference) don't provide immediate clinical context. Therefore, legitimate clinical importance can only be determined by framing the difference in means (± confidence intervals) with relevant Minimal Detectable Change (MDC) and Minimal Clinical Important Difference (MCID) thresholds. MDC represents 'the amount of change (in the outcome) that must be observed before it is considered above the bounds of measurement error'; and MCID represents 'the smallest change (in the outcome) that would be important to patients'. These thresholds are commonly overlooked, and a 2018 audit found that just 7% of orthopaedic researchers referred to MCID when determining treatment effects.[5]

Figure 1 shows data from two experimental studies,[6 7] each reporting statistically significant changes in ankle dorsiflexion post intervention ($p<0.05$). Despite this, the treatment effects observed in study A[6] cannot be differentiated from measurement error. Although study B[7] reports a larger average effect, most of the observed changes do not reach the threshold for clinical importance (MDC+MCID) and are unlikely to be meaningful to patients.

Insert Figure 1

Figure 1 footnote:
Dots (and whiskers) represent mean change scores (95% CIs)
Ankle dorsiflexion measured through a weight bearing lunge test (cm).; MDC = 1.9cm; MCID = 2cm

4. Replication

The replication crisis is a ubiquitous and complex problem across all of science. Sports medicine has been slower to react compared to other fields of medicine; currently, the volume of research in this field which has been successfully corroborated through replication is unclear. FAIR reminds clinicians and researchers that independent replication underpins scientific discovery; and that it is presumptuous to conclude treatment effectiveness based on a single significant result.

Summary

Time restraints and lack of training are cited as common barriers preventing clinicians from fully engaging in the evidence base. P-value thresholds (is $p<0.05$?) offer a fast but ultimately limited method for determining clinical effectiveness. Although there are many other aspects of trial design and reporting that can increase the risk of false discovery; FAIR is presented as a preliminary concept to help clinicians disentangle true positive from potentially false positive claims within sports medicine.

Competing interests: no competing interests for any author
Contributorship: Both authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. CB and JS were involved in the concept, design and writing. All authors were involved in final submission and revision of the manuscript.
Acknowledgements: none
Funding info: none
Ethical approval information: not applicable
Data sharing statement: not applicable

References

1. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2(8):e124. doi: 10.1371/journal.pmed.0020124
2. Colquhoun D. The False Positive Risk: A Proposal Concerning What to Do About *p*-Values. *The American Statistician* 2019;73:192-201.
3. Bleakley CM, Reijgers J, Smoliga JM. Many high quality RCTs in sports physical therapy are making false positive claims of treatment effect: a systematic survey. *J Orthop Sport Physio, in press*
4. Forstmeier W, Wagenmakers EJ, Parker TH. Detecting and avoiding likely false-positive findings - a practical guide. *Biol Rev Camb Philos Soc* 2017;92(4):1941-68. doi: 10.1111/brv.12315
5. Copay AG, Eyberg B, Chung AS, et al. Minimum Clinically Important Difference: Current Trends in the Orthopaedic Literature, Part II: Lower Extremity: A Systematic Review. *JBJS Rev* 2018;6(9):e2. doi: 10.2106/JBJS.RVW.17.00160
6. Driller M, Mackay K, Mills B, et al. Tissue flossing on ankle range of motion, jump and sprint performance: A follow-up study. *Phys Ther Sport* 2017;28:29-33. doi:10.1016/j.ptsp.2017.08.081
7. Marrón-Gómez D, Rodríguez-Fernández ÁL, Martín-Urrialde JA. The effect of two mobilization techniques on dorsiflexion in people with chronic ankle instability. Phys Ther Sport. 2015 Feb;16(1):10-5. doi: 10.1016/j.ptsp.2014.02.001

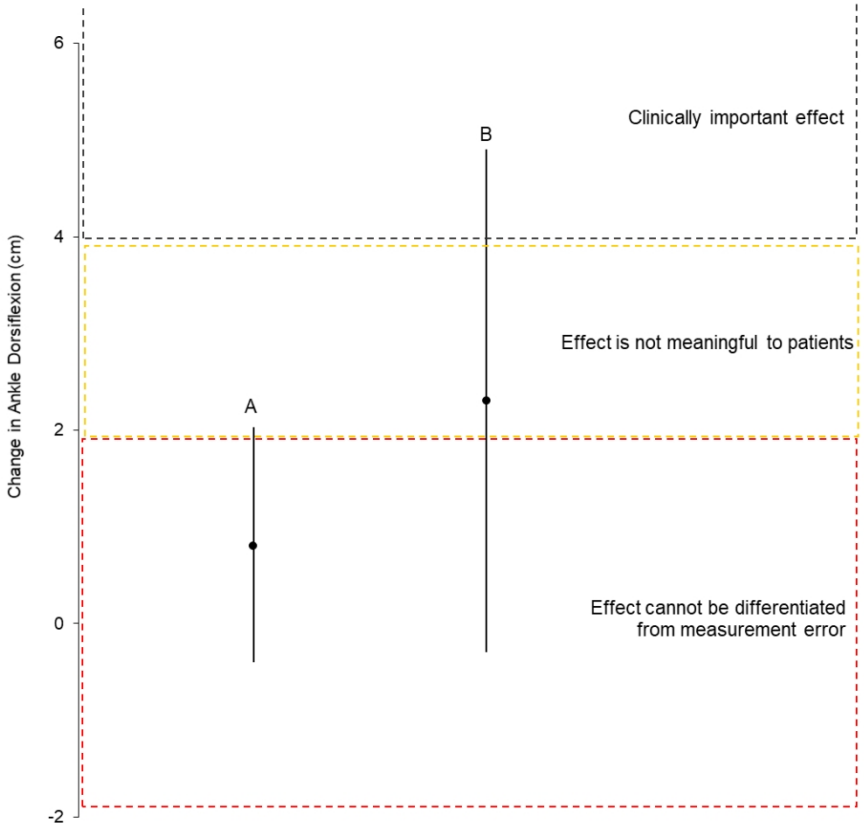

Figure 1 footnote:
Dots (and whiskers) represent mean change scores (95% CIs)
Ankle dorsiflexion measured through a weight bearing lunge test (cm); MDC = 1.9cm; MCID = 2cm